DOCUMENT RESUME

ED 389 725 TM 024 225

AUTHOR	Pomplun, Mark; And Others
TITLE	An Exploration of the Stability of Freshman GPA, 1978-1985.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-91-40
PUB DATE	Jun 91
NOTE	44p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Achievement; Achievement Tests; *College Freshmen; Educational Change; *Educational Trends; *Grade Point Average; High Schools; High School Students; History; Knowledge Level; Language Proficiency; Mathematical Aptitude; *Predictive Validity; *Reliability; Sciences; Verbal Ability; Writing Skills
IDENTIFIERS	Confirmatory Factor Analysis; *Scholastic Aptitude Test; *Test of Standard Written English

ABSTRACT

This study, one of several examining a decline in the predictive validity of the Scholastic Aptitude Test (SAT) and high school grades between 1975 and 1988, asks whether the criterion, freshman grade point average (FGPA), has changed. College Board tests usually thought of as predictors were used as proxies for the concurrent academic competencies that comprise FGPA. Academic components, defined by combinations of SAT, Test of Standard Written English (TSWE), and Achievement Test scores and high school grade point average, included verbal and mathematical reasoning ability, mathematics, language, and writing skills, and knowledge of science and history. Confirmatory factor analysis was used to test the consistency between 1978 and 1985 of the relationship between academic factors and FGPA. Primary results indicate that FGPA has been stable from 1978 to 1985 in relation to the abilities, skills, and subject knowledge measured by the SAT, TSWE, Achievement Tests, and high school grade point average. No evidence was found that a change in meaning of the FGPA has contributed to a decline in predictive validity of the SAT and high school grades. (Contains 14 tables, 2 figures, and 13 references.) (SLD)

RESEARCH REPORT

AN EXPLORATION OF THE STABILITY OF FRESHMAN GPA, 1978-1985

Mark Pomplun
Nancy Burton
Charles Lewis

Educational Testing Service
Princeton, New Jersey
June 1991

An Exploration of the Stability of Freshman GPA, 1978-1985

Mark Pomplun
Nancy Burton
Charles Lewis

ABSTRACT

This study is one of several meant to explain a decline in the predictive validity of the SAT and high school grades between 1975 and 1988. This particular study asks whether the criterion, freshman grade point average (FGPA), has changed. There may have been changes in the mix of academic competencies that students use in earning freshman grades. Possible changes in general reasoning abilities, technical skills, or subject knowledge were considered. College Board tests that are usually thought of as predictors were used in this study as proxies for the concurrent academic competencies that comprise FGPA. The academic components or factors, defined by various combinations of SAT, and TSWE items and scores, HSGPA, and Achievement Test scores, included verbal and mathematical reasoning ability; mathematics, language, and writing skills; and knowledge of science and history. Confirmatory factor analysis was used to test the consistency between 1978 and 1985 of the relationship of the academic factors to FGPA. The primary results indicate that FGPA has been stable from 1978 to 1985 in relation to the abilities, skills and subject knowledge measured by the SAT, TSWE, Achievement Tests, and HSGPA. No evidence was found to suggest that a change in the meaning of FGPA has contributed to the predictive validity decline.

This study was designed as an initial exploratory analysis, to be followed if warranted by a study with better proxies for the academic components of FGPA and with the addition of noncognitive variables such as study habits and interest. However, the results were so stable over a number of different proxies, and over a number of different ways of combining proxies, that the researchers conclude that no follow-up study is warranted.

An Exploration of the Stability of Freshman GPA, 1978-1985

This study is one of several seeking information about a decline in the predictive validity of the SAT and high school grades for a self-selected group of colleges (Willingham, Lewis, Morgan, and Ramist, 1990). These were colleges that elected to submit data to the College Board Validity Study Service for freshman classes entering between 1964 and 1988. (For a further description of these trends, see Morgan, 1989 and 1990; Ramist and Weiss, 1990). Most of these studies ask whether changes in the predictors or the populations involved might be related to the observed validity trends; this study addresses the question of whether the criterion, freshman grade point average (FGPA), may have changed.

We will look for possible changes over time in the mix of skills, abilities, and subject knowledge that students are required to use in earning their freshman year grades. The proxies for academic skills are those available in the Educational Testing Service files: the verbal and mathematical reasoning abilities measured by the SAT; the skills involved in responding to the Test of Standard Written English (TSWE) or the English Composition Achievement Test (ECT); the mathematics and foreign language reading skills measured by other Achievement Tests; and the subject-area knowledge covered in the history and science Achievement Tests.

These measures are necessarily somewhat narrow, in that they are based on one-hour multiple-choice tests[1] for which content specifications are established by consensus of secondary and college teachers from all regions of the country. Typically earned during the junior or senior year of high school, the test scores are also somewhat outdated as proxies for the skills utilized in earning FGPA. The academic skills that they estimate probably improve throughout high school and college, and at different rates for different students. Nevertheless, these test scores are valuable resources, in that they cover a wide spectrum of ability, knowledge, and skill, they are comparable across all students and colleges, and, most important, their specifications have remained almost completely constant over the period covered by this study.

The current study was designed to be an initial exploratory analysis. The purpose was to generate hypotheses, based on actual data, about possible changes over time in what academic competencies are useful in the freshman year of college.

The results of this study must be considered preliminary, both because the proxies were not collected concurrent with FGPA, and because they do not cover all the abilities (particularly the non-cognitive abilities) used in the freshman year. On the other hand, many important changes in the construct measured by FGPA should be at least somewhat reflected in these analyses. Any area of change detected in the current study design will require further investigation in a group of colleges willing to collect concurrent measures of skills, knowledge, and ability, as well as some important noncognitive components.

Because this is an exploratory study, there is no list of hypotheses to be tested. Instead, we defined a large number of different models, containing combinations of variables that might help capture how the academic part of the "job" of being a college freshman may have changed, and examined each model for evidence of change over time.

While the analysis appears similar in design to a conventional predictive validity study, with FGPA being regressed on a set of measures designed to predict success in college, the regression analysis is in fact being used to make inferences about whether the academic skills required in freshman classes have changed. The measures that would usually be thought of as "predictor variables" are being used instead as rough proxies of the concurrent academic components that make up the multidimensional composite FGPA. Possible changes over time in the components of FGPA are evaluated by controlling as many external sources of change as possible.

- Academic proxies were kept constant over time by using College Board tests that are developed to rigidly controlled content and statistical specifications and equated to a common scale. The only proxy that could not be so controlled was the high-school grade-point average (HSGPA). However, instead of using actual HSGPA, which was not available on all students, we used an average of self-reported grades in six academic subjects, which controlled for variations in curriculum among students and over time.
- To lessen the influence of chance variations in the proxy measures, academic factors were identified using LISREL. These factors should be less likely than observed scores to vary due to chance differences in measurement quality for the samples being analyzed.
- To control for differences in the composition of the sample of colleges at two time points, only colleges with data available for both of the entering classes being studied were included in the sample.

With these external sources of variation held constant, two entering freshman classes -- from the fall of 1978 and the fall of 1985 -- can be compared to determine whether the academic components of FGPA appear to have changed over time. Several examples of the kinds of changes that may have occurred follow.

The kinds of courses freshmen take may have changed. There has been a growth in the popularity of courses in technical areas, such as engineering and computer science (College Board, 1985), that might have led to a decline in the importance of the skills and abilities measured by the SAT-Verbal, the TSWE, and the English, history, and foreign language Achievement Tests.

There was also a trend through the '70s and early '80s for some colleges to drop or modify their core requirements; it may be that FGPA became less comparable from student to student and, therefore, less predictable. In that case, one would expect a decline in the contribution of all academic indicators.

Some colleges may have replaced academic core courses with more practical courses -- in business or vocational areas, for example -- that do not make heavy demands on analytical reasoning skills. In that case, one would predict a decline in the contribution of the more academic and analytical SAT, but not necessarily in specific skill or knowledge measures, such as the English or Mathematics Achievement Tests.

PROCEDURE

Sample

The data for this study were chosen from November 1977 and November 1984 administrations of the SAT and the TSWE, which included nearly half of the cohort of graduating seniors who took the SAT in those years[2]. The samples were 1978 and 1985 high school graduates who had SAT and TSWE item response data from those administrations. These files were matched with College Board Validity Study Service (VSS) records to obtain freshman grades for the subset who attended participating colleges.

In all, 52,948 records were selected from the 1985 group and 6,814 from the 1978 sample. (The sample is smaller for 1978 because only a spaced sample of one-seventh of the data was retained in ETS program archives.) The students were selected using the following criteria: that English was the best language, complete SAT and HSGPA data were available, and FGPA was reported on a four-point scale. The students were divided into two subsamples: One consisted of students from colleges that did validity studies for both the 1978 and 1985 entering freshmen and the other contained the students from nonmatched colleges. The matched sample, which included 86 colleges with 2,984 students in 1978 and 32,338 students in 1985, was used to analyze the stability of FGPA over time. The nonmatched sample was used for exploratory model fitting.

Because of institutional changes over the years in areas like remedial courses and Advanced Placement courses (College Board, 1987; College Board, 1988), it is possible that the validity trends will not be the same for different ability groups. If this study detects changes to FGPA over the years studied, then these changes will also be studied at different ability levels when possible. For analyses with samples over 400, the examinees were also divided into three ability groups. A weighted combination of the SAT-V, SAT-M, and high school grade point average (HSGPA), developed by Ramist, Lewis and McCamley (1990), was used to divide the sample into ability groups. A division into groups of about 27%, 46% and 27% of the cases was suggested by Angoff, Pomplun, McHale, and Morgan (1990).

Measures and Models

Although measures of various abilities (SAT-Verbal and SAT-Mathematical); skills (such as writing and foreign language fluency); subject areas (such as knowledge of science or history) and grades were available, these are observed measures with measurement error. In this study models were developed using subscores from the SAT and TSWE to estimate factors free from measurement error that were then correlated with observed FGPA.

These models were analyzed through LISREL VI (Joreskog and Sorbom, 1986), which allows the user to combine factor analytic and regression models into one. The factor analytic part of the model specifies how observed variables are related to the latent factors; the regression part specifies the relationship between the latent factors and the criterion factor. In this study, the model's factor analytic part defines various latent academic factors, such as verbal and mathematical reasoning and writing skill, through SAT and TSWE subscores; the regression part specifies the relationship between the latent academic factors and the observed criterion, FGPA.

A group of SAT-TSWE models were based on data that were available for the entire sample -- high school grades and measures that could be constructed from SAT and TSWE items. SAT items were divided into subscores based on item format (analogies, quantitative comparisons, etc.), content (knowledge of vocabulary in science, algebra, geometry, etc.), or cognitive level (application, insight, etc.) as defined by Rock especially for this study (D. Rock, personal communications, October 10 and 20, 1989). TSWE items were divided into item format subscores only, because each of the numerous content classes contained too few items to form subscores that could be compared over time. Because multiple measures of HSGPA and FGPA were not available, it was not possible to estimate latent factors for these two variables.

In the item type model (Table 1), SAT-V subscores created from the analogies, antonyms, reading comprehension, and sentence completion item formats are the observed measures defining a verbal reasoning factor. SAT-M subscores from the quantitative comparison and regular multiple-choice item formats define the mathematical reasoning factor. TSWE subscores based on the sentence correction and usage formats define the writing skill factor. Student-reported HSGPA was the only available measure of high school achievement. Freshmen success as measured by FGPA was regressed on the verbal, mathematical, writing and HSGPA factors.

The item type model can provide tentative information on several hypotheses about freshman grades. For example, an increase in the contribution of reading comprehension to the prediction of FGPA could mean that there has been an increase in the amount or difficulty of reading required of freshmen. For another example, a decrease in the importance of the some item types (verbal analogies or mathematical quantitative comparisons) may mean that the more abstract reasoning abilities are now less in demand. Figure 1 displays the formal representation of the Item Type Model. (The parameter estimates are explained on page 15.) Table 1 defines the abbreviations used in Figure 1.

Insert Table 1 and Figure 1

In the item content model (Table 2), verbal and mathematical reasoning are defined by subscores constructed for the purpose of this study based on item-content specifications for the test. SAT-V practical affairs, human relations, science, and aesthetic/philosophical subscores define the verbal reasoning factor. (Because the content categories of reading comprehension items changed across the two years, and because the test development specialist could find no good way to collapse the two categorization systems, the reading items were not included in the item-content analysis.) SAT-M geometry, algebra, arithmetic, and miscellaneous mathematics subscores define the mathematical reasoning factor. Freshmen success as measured by FGPA was regressed on the verbal, mathematical, writing and HSGPA factors. The TSWE and HSGPA measures are the same as in the previous model.

The item content model, like the Achievement Test models, can provide tentative information about possible changes in the contribution of more specific knowledge and skills to freshman grades. For example, an increase in the number of practical and technical freshman courses might lead to an increase in the importance of "practical affairs" vocabulary, science vocabulary, or arithmetic problem-solving skills.

Insert Table 2

In the item cognitive level model (Table 3), three cognitive-level subscores were defined using only the SAT reading comprehension items (D. Rock, personal communication, October 20, 1989). It was possible to use all the mathematical items in defining the math cognitive levels (D. Rock, personal communication, October 10, 1989). These subscores, unlike the item-type and item-content subscores, were not based on the specifications used to develop the tests.

Three SAT-V cognitive-level subscores (reproduction, evaluation, and inference) define the verbal factor and three SAT-M cognitive-level subscores (application, insight, and production) define the mathematical factor. The ***reproduction*** subscore measures the ability to reproduce details from the reading passage. ***Evaluation*** requires an interpretation of the author's main thought. The ***inference*** level represents the ability to go beyond the author's main thought, as when making an inference. The first mathematical cognitive level ability is *application*, which measures knowledge with respect to the rote application of various simple arithmetical and algebraic rules. *Insight* involves the rote application of simple computational rules plus an additional insightful or logical step. The ***production*** level first requires the translation of a logically complex verbal statement into the proper equation. After the translation, the correct solution is obtained with another step following procedures characteristic of the application level.

If freshman classes have been influenced by the influx of less prepared students, one might expect the lower cognitive levels, reproduction in verbal areas and application in mathematical areas, to grow more important in the definition of their factors over time.

Insert Table 3

The item type model is the most inclusive of the SAT-TSWE models because the verbal reasoning factor in the item content model does not utilize reading comprehension items; only the reading comprehension items are used to define the verbal factor in the item cognitive level model. The SAT specifications call for the item type subscores to be comparable in number of items and mean and standard deviation of difficulty, and consequently, should be more parallel over time than the other sets of subscores included in this study.

The second group of models was a series that included Achievement Test scores along with SAT and HSGPA measures. Because only about one SAT examinee in every five takes any Achievement Test, (College Board, 1978, 1985) the samples available for these analyses were much smaller. Ideally, all Achievement Tests would be included in the same regression model to estimate changes over time in the mix of subject skills that influence FGPA, but the combination of small 1978 samples and missing correlations among Achievement Tests (most students who take these tests take about 3 of the 15 (College Board, 1985)) made several compromises necessary in the analysis of models that include Achievement Test measures.

One result of the small sample sizes was the impossibility of forming low, middle and high ability groups for the Achievement models. The small samples also compelled us to take advantage of the fact that all Achievement Tests are placed on a common scale. In order to enlarge samples used in the analyses, a score on any of the science tests (Biology, Chemistry, and Physics) was treated as a Science score. Scores on all foreign language tests (Spanish, French, Latin, German, Hebrew, and Russian) were also treated as interchangeable, as were scores on Mathematics Levels I and II (as Mathematics scores) and American and European History (as History scores).

There was a separate model for each achievement area. Each model included one achievement area and also included the SAT item type subscores, as measures of the latent verbal and mathematical reasoning factors, and HSGPA. The SAT item type subscores were used because they were the most stable of the SAT subscores across years in the exploratory stage of the study. In each achievement model, FGPA was regressed on the verbal, mathematical, HSGPA and achievement factors. The achievement models are shown in Tables 4 through 8. Two achievement models, Science and Mathematics (Tables 7 and 8), did not include the SAT mathematical reasoning factor because of collinearity between the scores on the mathematical reasoning factor and the Science and the Mathematics Achievement area scores. Figure 2 portrays an Achievement model from Table 5. (The parameter estimates are explained on page 15.)

Insert Tables 4 - 8 and Figure 2

Analysis

The variables chosen for analysis were standardized within each college before all students were pooled into a single analysis sample. The regression equation derived from the models can be considered to be a weighted average of within-college equations.

LISREL was used to assess the fit of the same overall model to the freshmen entering college in 1978 as compared to those entering in 1985. In addition to looking at the overall fit, we examined specific parameters relating the observed proxies to the academic factors (the "factor analytic" part of the model) and the parameters relating the academic factors to the observed FGPA (the "regression" part of the model).

The final models were selected from several alternatives evaluated using the data from colleges participating in only one of the two years. As mentioned previously, some collinearity between scores for the SAT factors and the Achievement Tests was solved by not including the mathematical reasoning factor created from SAT-M subscores in the final Science Achievement and Mathematics Achievement models.

The models arrived at during the exploratory stage were tested in the matched college data set. The LISREL system allows the data analyst to constrain in various ways all possible pairwise relationships between the study variables, both latent and observed. The research questions were investigated by comparing the results for two different levels of constraint on each model.

A condition assuming that the factor loadings[3] and regression weights for the two years were similar was contrasted for the same model with a condition assuming that these values were identical for the two years. The least restricted condition studied is called the "similar" condition. In this condition the pattern of relationships between the observed variables and the factors were set to be similar but not identical in the two years. "Similar" is defined as follows: the size of the factor loadings and regression weights for the variables assumed to be related was allowed to vary but for the variables assumed to be unrelated, parameters were set to zero in both years.

In the more restricted condition, called "invariant", all factor loadings and regression weights were required to be equal in 1978 and 1985. A substantial decrease in the fit of the more restrictive condition compared to the "similar" condition can be interpreted as evidence that there has been a change over time in the appropriate model.

The multisample analysis feature of LISREL allows a variety of tests of the fit of a single model to two different data sets. LISREL combines the two data sets being compared into a weighted total data set. In interpreting the fit results it is important to recall that the samples for 1985 are more than 10 times larger than those for 1978. Thus data from 1985 dominate when the two years are combined and combined statistics will very closely resemble the 1985 values. It may be helpful to think of the fit indices as measuring the extent to which the 1978 freshmen can be treated as a sample of the "population" defined by the larger group of 1985 freshmen.

Although several measures of overall fit are available, this study used the goodness-of-fit index provided by LISREL and an adaptation of the Tucker-Lewis index (Tucker & Lewis, 1973) to assess the fit of the models. Fit for each year separately was assessed by the LISREL goodness-of-fit index and fit for both years together by the ***Tucker-Lewis partial index***. The goodness-of-fit index measures how well the model reproduces the matrix of variances and covariances. This index was developed by Joreskog and Sorbom for Maximum

Likelihood estimation and can be written as follows:

$$GFI_{ML} = 1 - tr\ (\Sigma^{-1}\ S-I)^2/tr(\Sigma^{-1}\ S)^2$$

where S is the sample covariance matrix and Σ is the reproduced covariance matrix based on the specified model. Joreskog and Sorbom (1986) considered it relatively robust to departures from normality. The index ranges between 0 and 1; for this study values above .90 are required for tentative acceptance of the fit of the model to the data.

The original Tucker-Lewis index represented the ratio of the amount of variance associated with a given model to the total variance. The adapted measure created for this study is an index that represents the percent of residual variance from the fit of a null model that is explained by an alternative model. Because it is based on residual variance rather than total variance, it can be viewed as analogous to a partial correlation. The Tucker-Lewis partial index is calculated in the following manner:

$$TLP = \frac{(X^2_N/df_N) - (X^2_A/df_A)}{(X^2_N/df_N-1)}$$

where X^2_N and df_N are the chi square and degrees of freedom associated with the null model and X^2_A and df_A are the same quantities associated with the alternative model. This index has an expected value of zero but is permitted to have small negative values. Whenever the index is close to zero, positive or negative, one can conclude that the same model and parameters fit the 1978 and 1985 data. The adaptation used in this study uses the invariant model as the null model and the less-restrictive similar model as the alternative model.

Examples of the full LISREL models with the paths and parameter estimates are displayed in Figures 1 and 2 above. In both Figures, the first number on the path is the parameter estimate for 1978 and the second for 1985. The squares represent observed variables and the circles latent variables. The paths from the latent variables to the observed show the factor loadings of the observed variable on the latent variable. The other path to the observed variables displays the residual term associated with the observed variable. The paths between the latent variables represent the regression of FGPA on the latent variables in the model. The remaining path to FGPA is the residual from the regression of FGPA on the latent predictors. It should be noted that when a latent variable is defined by only one observed variable, as is the case here for FGPA and HSGPA, the factor loading is 1.00 and the residual 0.00 for the observed variable. Thus, while the verbal, mathematical and writing factors are estimates of true scores, the HSGPA and FGPA factors are actually observed score variables containing measurement error.

RESULTS

The analysis focuses on changes between the 1978 and 1985 entering classes. The fit of the models for each class and the fit of a single model to both entering classes are compared. The most detailed analyses evaluate the various parameters of the regression model. An attempt is made to generate hypotheses about the most likely location of any observed changes -- whether they appear to be occurring mainly in the latent factors being measured or in how those measures contribute to the prediction of freshman success.

Table 9 displays the fit indices for the SAT-TSWE and achievement models for the two years. The goodness-of-fit indices are high for all models in both conditions. For all variations, the models reproduce almost all of the observed variances and covariances.

Insert Table 9

Table 9 also displays the Tucker-Lewis partial index which compares the similar and invariant conditions. The index represents the improvement in fit obtained by allowing the size of factor loadings and regression weights to vary over the two years. For all model variants the Tucker-Lewis partial index indicates that relaxing the model assumptions does not improve the fit. Both fit indices agree that the models and their parameters appear nearly identical for the two years. In other words, Table 9 shows no evidence of change over time.

Because the primary concern of this study was change in the academic components of FGPA, the parameters for the regression of FGPA on the latent variables are shown in Tables 10 and 11. Table 10 displays the regression weights and standard errors for the SAT-TSWE models and Table 11 displays the regression weights and standard errors for the achievement models. The achievement area of History showed a significant decrease of 2.6 standard errors based on the larger of the two standard error estimates ((.265 - .101)/.064). When standard errors are considered, none of the other changes in regression weights appears statistically significant in either table. Thus the regression weights displayed are also consistent with the results from the fit indices. There is little or no evidence of change in FGPA between 1978 and 1985.

Insert Tables 10 & 11

We were also concerned secondarily with possible trends in predictive validity over time. These trends can best be evaluated by inspecting the coefficients of determination for each model and entering class. This index represents the proportion of the variance in the criterion that is explained by the model. It is analogous to the squared multiple correlation between the predictors and the criterion corrected for unreliability in the predictors. Table 12 displays the coefficients of determination for the three SAT-TSWE models. The coefficient of determination can be contrasted with the R^2 index, also reported in Table 12, from a multiple regression of the observed criterion on the observed predictors. The R-Squared is presented as a benchmark for how well the models represent the observed relationships in the data.

The slight decline in validity over time for the total group, displayed by the coefficient of determination and R-Squared for all three SAT-TSWE models, mirrors the results reported by other researchers (Morgan, 1989, 1990; Ramist & Weiss, 1990). Table 13 displays the coefficients of determination and R-Squares for the analyses of the different achievement models. The results include nonsignificant increases over the years for the Foreign Language model.

The results of the analysis of indices of goodness-of-fit differ from the results of the analysis of coefficients of determination. The goodness-of-fit analyses all showed no change over time in the relations among academic proxies, academic factors, and FGPA, while the secondary analyses shows a decline over time in the proportion of total variance accounted for by the model. These results are not inconsistent. The secondary analyses compared the explained variance (the model) to total variance, while the primary analyses were concerned with the relationships among factors in the model.

In other words, the results of our analysis suggest that the decline that other researchers have observed in the predictive validity of admissions measures, including both high school grades and test scores, is due to an increase in residual or unexplained variance rather than to a change in the academic skills and abilities rewarded in the first year of college.

Insert Tables 12 and 13

Table 14 further investigates the SAT-TSWE model validity trend by displaying the coefficients of determination for 1978 and 1985 for the item type model by ability group. It should be noted that the interpretation of ability levels is for a within-college analysis. Because all variables were standardized within college in the pooled within-college analysis, the high-ability students (for example) are not the highest 27% of all students studied, but rather the highest 27% from each college. Thus the within-college ability-level analysis concentrates on validity by relative standing on the combined predictors: Has predictive validity changed for those in the top, middle, or bottom of their respective entering college classes?

The coefficients of determination for the item type model showed a small increase for the low and medium ability groups. The high ability group was the only group to display a slight decrease from 1978 to 1985. However, despite the decline in the coefficients for the high ability group, that group still showed the highest percent of variance accounted for in both years.

Insert Table 14

In contrast, Morgan (1990) found validity declines mostly for low and middle ability students and only slight declines for high ability student and Ramist, Lewis and McCamley (1990) observed a validity increase for the high ability group and a validity decrease for the low ability group. These inconsistencies could be due to different sampling and analysis strategies employed by the different studies.

For example, Morgan (1990) examined predictive-validity trends for matched pairs of colleges participating in the College Board Validity Study Service (VSS) in the years 1978, 1981, and 1985. While a number of the same colleges and the same data points are included in Morgan's study and our own, there are numerous differences in detail between the two studies and the two samples. Because our design required that all students take the same test edition, we confined ourselves to scores from the largest administration in each year studied while Morgan analyzed all available scores. Because Morgan did not pool his data, he confined himself to the subset of colleges with 25 or more freshmen in each year, while we required only 7 freshmen; he used least squares multiple linear regression of observed predictors of FGPA rather that a LISREL analysis; he corrected for restriction of range on SAT scores and HSGPA while we did not. Finally, Morgan defined the ability levels based on a single predictor at a time and divided the groups into equal thirds. The discrepancy in results between this study and Morgan's and Ramist et al. emphasizes the need for great caution in interpreting the results of such complex analyses on self-selected samples.

In fact, the overall validity trend being analyzed -- a decline in the average college correlation coefficient of about .04 points -- is extremely small when compared to total variance. As Willingham notes, (Willingham et al., 1990, p. 32), the linear trend in validity coefficients for the SAT and HSGPA accounts for only 2% of total variance, while other systematic differences among colleges account for 77% of total variance. Thus the results of any such study are likely to depend heavily on the exact sample of colleges and years studied.

CONCLUSIONS

The purpose of this study was to investigate possible changes in FGPA as a factor in the decline of SAT validity. The changes investigated were primarily in either the mix or the level of several academic competencies -- general analytic ability, technical skills, and subject knowledge. For this initial exploratory analysis, the academic factors were measured by the convenient, if rough, proxies formed by the rescoring the SAT and the TSWE, and using also total Achievement Test scores and HSGPA. The primary finding is that the academic proxies have a stable relationship with FGPA. The 1978 SAT form has the same factor structure as the 1985 SAT form, and the academic factors and FGPA have the same regression structure in 1978 and 1985. The results from the analysis of the Achievement models supported this conclusion, despite a suggestion that history knowledge may have become less important by 1985. Despite the limitations in the proxies used to define the dimensions of FGPA, the fact that results stayed stable over a number of different proxies, and over a number of different ways of combining proxies, suggests that FGPA has not changed in the colleges included in this study. This study found no evidence that a change in the meaning of FGPA has contributed to the predictive validity decline described in Willingham et al. (1990).

The percent of total FGPA variance explained by all three SAT-TSWE models decreased from 1978 to 1985, in accordance with other analyses of validity trends. Because the same SAT-TSWE models fit both years the validity decrease could not be connected with any specific area of the tests of HSGPA. Instead, the decline appears to be a result of an increase in residual, or unexplained, variance over time.

When the validity decline was examined at the ability group level, the results were inconsistent with other studies of the same decline (Morgan, 1990; Ramist, Lewis & McCamley, 1990). Despite using similar data bases and having related goals, the results of all three studies were inconsistent with each other, most likely because of rather small differences in the sampling and analysis strategies used in the different studies. The inccnsistencies point out the need for caution when drawing conclusions about the validity decline.

In summary. the results suggest that the academic competencies used in earning FGPA have been stable over the years at least in so far as the SAT, TSWE, Achievement Tests, and HSGPA serve as proxies for academic competency. Verbal and mathematical reasoning abilities, writing and language skills, and subject area knowledge in history and science continue to be important components of academic performance.

Our recommendation, based on this exploratory study, is not to pursue a more complex study of the academic components of FGPA. Remarkably little evidence of any type of change was found. On the other hand, other studies (Willingham et al., 1990) have developed a number of hypotheses having to do with institutional characteristics that do seem to be related to the observed declines in validity coefficients.

NOTES

1. Once a year, the English Composition Test includes a 20-minute impromptu essay; about 40% of the students submitting ECT scores take this version of the test.

2. The years 1977 and 1984 were chosen as the earliest and latest years that fit two criteria -- a large number of validity studies done and item data available on file.

3. Here the weights on the latent variables for reproducing the observed variables are labeled factor loadings, to easily distinguish them from the weights on the latent variables reproducing FGPA, labeled regression weights.

References

Angoff, W. H., Pomplun, M., McHale, F. & Morgan, R. (1990). Comparative study of factors related to the validities of 1974-75 and 1985-85 forms of the SAT. In W. Willingham, C. Lewis, R. Morgan and L. Ramist (Eds.), Predicting college grades: An analysis of institutional trends over two decades. Princeton, NJ: Educational Testing Service.

College Board (1978). National college-bound seniors, 1978. New York: College Board.

College Board (1985). National college-bound seniors, 1985. New York: College Board.

College Board (1987). Annual survey of colleges, 1986-87: Summary statistics. New York: College Entrance Examination Board.

College Board (1988). The College Board technical manual for the Advanced Placement Program. New York: College Entrance Examination Board.

Joreskog, K.G. & Sorbom, D. (1986). LISREL VI: Analysis of linear structural relationships by the method of maximum likelihood. Mooresville, IN: Scientific Software.

Morgan, R. (1989). Analyses of the predictive validity of the SAT and high school grades from 1976 to 1984. (College Board Report No. 89-7 and ETS Research Report No. 89-37). New York: College Board.

Morgan, R. (1990). Predictive validity within categorizations of college students: 1978, 1981, and 1985. In W. Willingham, C. Lewis, R. Morgan and L. Ramist (Eds.), Predicting college grades: An analysis of institutional trends over two decades. Princeton, NJ: Educational Testing Service.

Ramist, L. (1984). Predictive validity of the ATP tests. In T. F. Donlon, (Ed.), The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests. New York: College Board.

Ramist, L., Lewis, C., & McCamley, L. (1990). The effects of freshman GPA on the observed predictive validity of the SAT and high school grade-point average, 1982 and 1985. In W. Willingham, C. Lewis, R. Morgan and L. Ramist (Eds.), Predicting college grades: An analysis of institutional trends over two decades. Princeton, NJ: Educational Testing Service.

Ramist, L. & Weiss, G. (1990). The predictive validity of the SAT, 1964 to 1988. In W. Willingham, C. Lewis, R. Morgan and L. Ramist (Eds.), Predicting college grades: An analysis of institutional trends over two decades. Princeton, NJ: Educational Testing Service.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Willingham, W. W., Lewis, C., Morgan, R. & Ramist, L. (1990). Predicting college grades: An analysis of institutional trends over two decades. Princeton, NJ: Educational Testing Service.

Table 1

Definition of Academic Factors:

Item Type Model

Observed Variable or Subscore	Abbreviation[a]	Academic Factor	Abbreviation[a]
SAT-V			
Antonyms	VAT	Verbal Reasoning	V
Analogies	VAN		
Sentence Completion	VSC		
Reading Comprehension	VRC		
SAT-M			
Regular Multiple Choice	MMC	Mathematical Reasoning	M
Quantitative Comparison	MQC		
Test of Standard Written English			
Sentence Correction	TSC	Recognition of Standard English	W
Usage	TUS		
High School Grade Point Average (self-reported)	HGP	Success in High School Courses	H
Freshman Grade Point Average	FGP	Success in Freshman Courses	F

[a] Abbreviation used in Figure 1.

Table 2

Definition of Academic Factors:

Item Content Model

Observed Variable or Subscore	Academic Factor
SAT-V	
Practical Affairs	Verbal Reasoning
Human Relations	
Science	
Aesthetic/Philosophical	
SAT-M	
Arithmetic	Mathematical Reasoning
Algebra	
Geometry	
Miscellaneous	
Test of Standard Written English	
Sentence Correction	Recognition of Standard English
Usage	
High School Grade Point Average (self-reported)	Success in High School Courses
Freshman Grade Point Average	Success in Freshman Courses

Table 3

Definition of Academic Factors:

Item Cognitive Level Model

Observed Variable or Subscore	Academic Factor
SAT-V	
Reproduction Evaluation Inference	Verbal Reasoning
SAT-M	
Application Insight Production	Mathematical Reasoning
Test of Standard Written English	
Sentence Correction Usage	Recognition of Standard English
High School Grade Point Average (self-reported)	Success in High School Courses
Freshman Grade Point Average	Success in Freshman Courses

Table 4

Definition of Academic Factors:

English Composition Model

Observed Variable or Subscore	Academic Factor
SAT-V	
Antonyms	Verbal Reasoning
Analogies	
Sentence Completion	
Reading Comprehension	
SAT-M	
Regular Multiple Choice	Mathematical Reasoning
Quantitative Comparison	
English Composition Test	Recognition/ Production of Standard English
High School Grade Point Average (self-reported)	Success in High School Courses
Freshman Grade Point Average	Success in Freshman Courses

Table 5

Definition of Academic Factors:

Foreign Language Skill Model.

Observed Variable or Subscore	Abbreviation[a]	Academic Factor	Abbreviation[a]
SAT-V			
Antonyms	VAT	Verbal Reasoning	V
Analogies	[illegible]		
Sentence Completion	vSC		
Reading Comprehension	VRC		
SAT-M			
Regular Multiple Choice	MMC	Mathematical Reasoning	M
Quantitative Comparison	MQC		
Foreign Language			
Achievement Tests[b] French or German or Hebrew or Latin or Spanish or Russian	FL	Grammar, Usage and Reading Skills in Foreign Languages	FL
High School Grade-Point Average (self-reported)	HGP	Success in High School Courses	H
Freshman Grade-Point Average	FGP	Success in Freshman Courses	F

[a] Abbreviations used in Figure 2.

[b] Scores on any of the six foreign language Achievement Tests treated as interchangeable.

Table 6

Definition of Academic Factors:

History Knowledge Model.

Observed Variable or Subscore	Academic Factor
SAT-V	
Antonyms	Verbal Reasoning
Analogies	
Sentence Completion	
Reading Comprehension	
SAT-M	
Regular Multiple Choice	Mathematical Reasoning
Quantitative Comparison	
History Achievement Tests[a]	
American History and Social Studies	
or	
European History and World Cultures	Knowledge of History
High school grade point average (self-reported)	Success in High School Courses
Freshman grade point average	Success in Freshman Courses

[a] Score on either of the two history Achievement Tests treated as interchangeable.

Table 7

Definition of Academic Factors:

Mathematics Skill Model.

Observed Variable or Subscore	Academic Factor
SAT-V Antonyms Analogies Sentence Completion Reading Comprehension	Verbal Reasoning
Mathematics Achievement Tests[a] Mathematics Level I or Mathematics Level II	Mathematics Knowledge and Skill
High School Grade Point Average (self-reported)	Success in High School Courses
Freshman Grade Point Average	Success in Freshman Courses

[a] Scores on either of the two mathematics Achievement Tests treated as interchangeable.

Table 8

Definition of Academic Factors:

Science Knowledge Model.

Observed Variable or Subscore	Academic Factor
SAT-V	
Antonyms	Verbal Reasoning
Analogies	
Sentence Completion	
Reading Comprehension	
Science Achievement Tests[a]	
Biology	Science Knowledge
or	
Chemistry	
or	
Physics	
High School Grade Point Average (self-reported)	Success in High School Courses
Freshman Grade Point Average	Success in Freshman Courses

[a] Scores on any of the three science Achievement Tests treated as interchangeable.

Table 9

Fit Indices for SAT-TSWE and Achievement Models

	Goodness of Fit				
	Restrictiveness of Model Assumptions				
	Similar		Invariant		
	1978	1985	1978	1985	Tucker-Lewis Partial Index
Model					
SAT-TSWE					
Item Type	.994	.991	.993	.991	-.17
Item Content	.991	.997	.989	.997	-.10
Item Cognitive Level	.995	.995	.988	.995	-.02
Achievement					
English Composition Test (ECT)	.993	.988	.991	.988	-.19
Foreign Language Area (FL)	.986	.987	.984	.987	-.22
History Area	.987	.982	.979	.982	-.17
Mathematics Area	.996	.989	.994	.989	-.27
Science Area	.992	.992	.988	.992	-.26

Table 10

Regression Weights and Standard Errors

SAT-TSWE Models

	Year	Writing	Verbal	Mathematical	HSGPA
Item Type	1978	.120 (.049)	.115 (.056)	.166 (.029)	.316 (.017)
	1985	.102 (.015)	.154 (.015)	.154 (.009)	.308 (.005)
Item Content	1978	.137 (.047)	.092 (.054)	.169 (.030)	.318 (.017)
	1985	.133 (.014)	.109 (.014)	.162 (.009)	.308 (.005)
Item Cognitive Level	1978	.137 (.047)	.130 (.061)	.170 (.031)	.315 (.017)
	1985	.124 (.013)	.171 (.015)	.177 (.012)	.308 (.005)

Table 11

Regression Weights and Standard Errors

Achievement Models

	Year	Achievement Test	Verbal	Mathematical	HSGPA
English Composition Test (ECT)	1978	.168 (.039)	.039 (.064)	.190 (.047)	.280 (.027)
	1985	.114 (.012)	.138 (.018)	.163 (.014)	.288 (.008)
Foreign Lang. (FL)	1978	.075 (.057)	.197 (.108)	.156 (.107)	.265 (.055)
	1985	.111 (.015)	.239 (.026)	.175 (.026)	.287 (.015)
History Area	1978	.265 (.064)	.089 (.109)	.027 (.088)	.236 (.054)
	1985	.101 (.026)	.171 (.041)	.130 (.031)	.297 (.018)
Mathematics[a] Area	1978	.201 (.028)	.225 (.040)	[a]	.272 (.027)
	1985	.150 (.009)	.246 (.013)		.287 (.008)
Science[a] Area	1978	.278 (.048)	.107 (.072)	[a]	.270 (.042)
	1985	.188 (.015)	.181 (.022)		.311 (.012)

[a] SAT-M was dropped from these Achievement Models because of collinearity.

Table 12

Coefficients of Determination (COD) and R-Squared[a] for SAT Models

SAT/TSWE Model	Year	N	COD	R-Squared
Item Type	1978	2,984	.240	.234
	1985	32,338	.222	.218
Item Content	1978	2,984	.238	.235
	1985	32,338	.219	.214
Item Cognitive	1978	2,984	.238	.234
	1985	32,338	.221	.215

[a] Based on observed-variable least squares regression.

Table 13

Coefficients of Determination (COD) and R-Squared[a] for Achievement Models

Achievement Test Model	Year	N	COD	R-Squared[a]
English Composition Test (ECT)	1978	1188	.214	.213
	1985	12999	.204	.204
Foreign Language Area (FL)	1978	308	.197	.192
	1985	3966	.229	.226
History Area	1978	327	.207	.213
	1985	2587	.200	.207
Mathematics Area	1978	1188	.219	.218
	1985	12999	.202	.199
Science Area	1978	475	.205	.208
	1985	5558	.216	.216

[a] Based on observed-variable least squares regression.

Table 14

Coefficients of Determination (COD) for Item Type model prediction of FGPA by Ability Groups

Ability Group	Year	N	COD
High	1978	801	.187
	1985	8767	.173
Medium	1978	1461	.089
	1985	15140	.110
Low	1978	722	.073
	1985	8431	.100

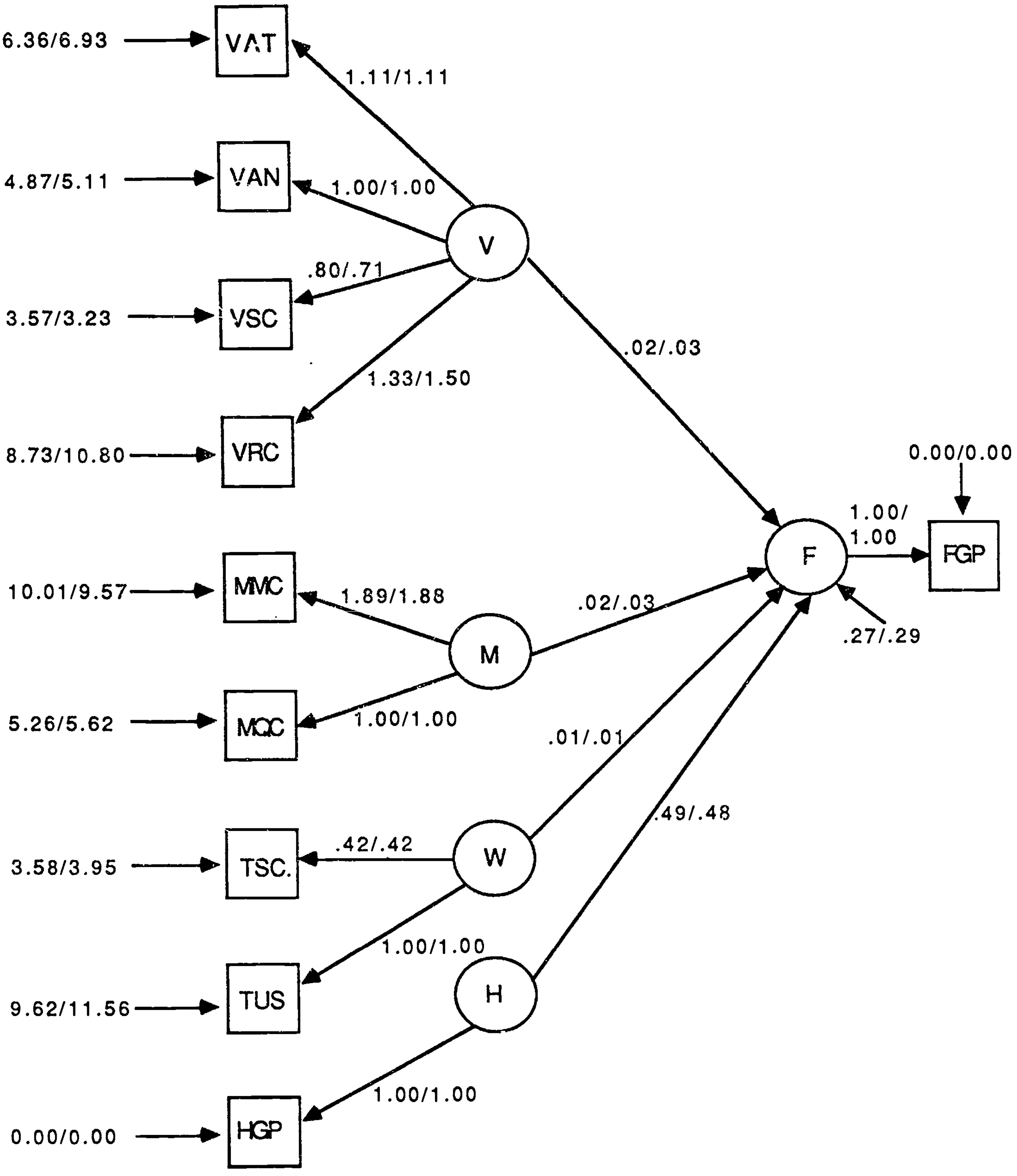

FIGURE 1. An example of the SAT-TSWE Models. The Item Types model (1978/1985).

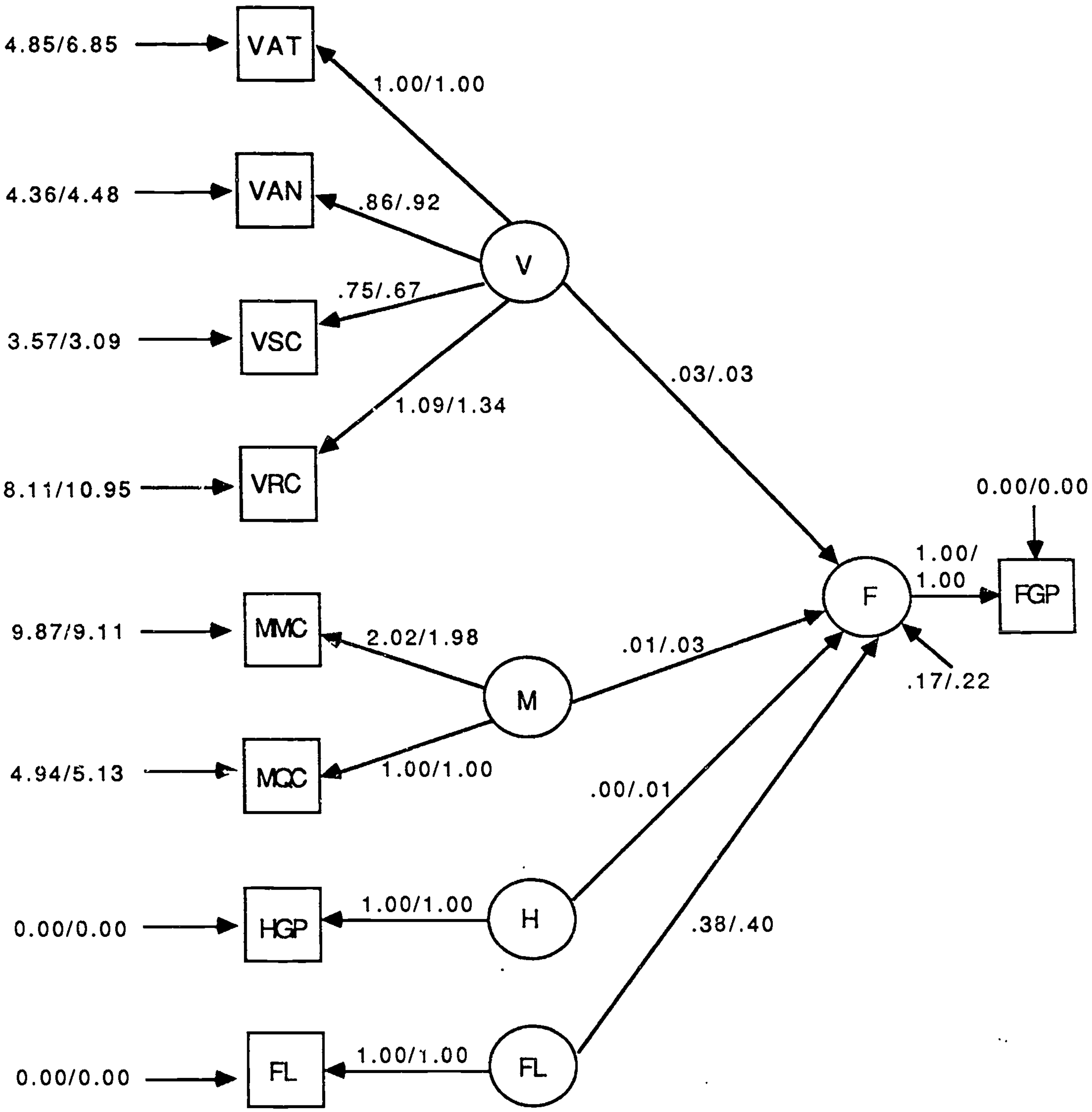

FIGURE 2. An example of the Achievement models. The Foreign Language Achievement model (1978/1985).